

Weekly Report

2013.12.02 – 2013.12.08

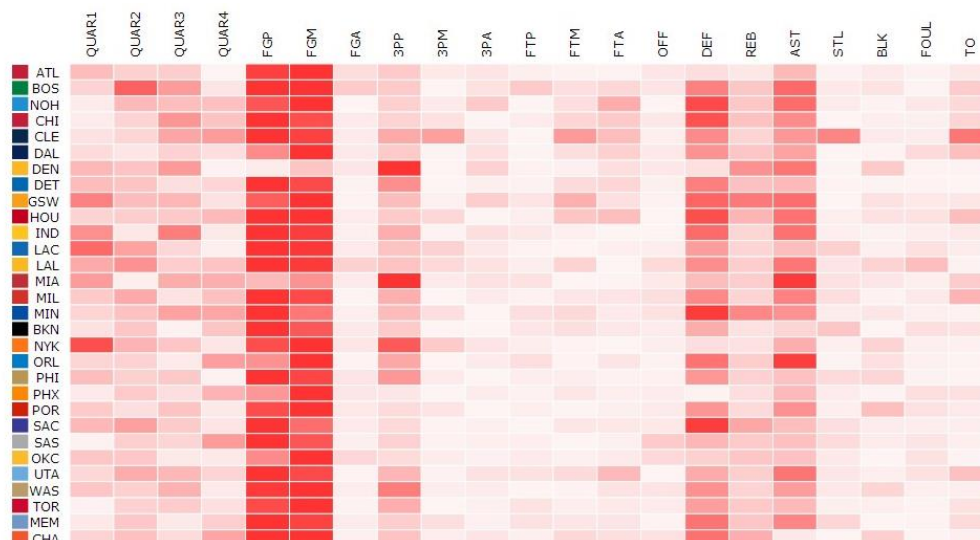
黄芯芯

本周工作：

1. NBA 赛事可视化项目报告:

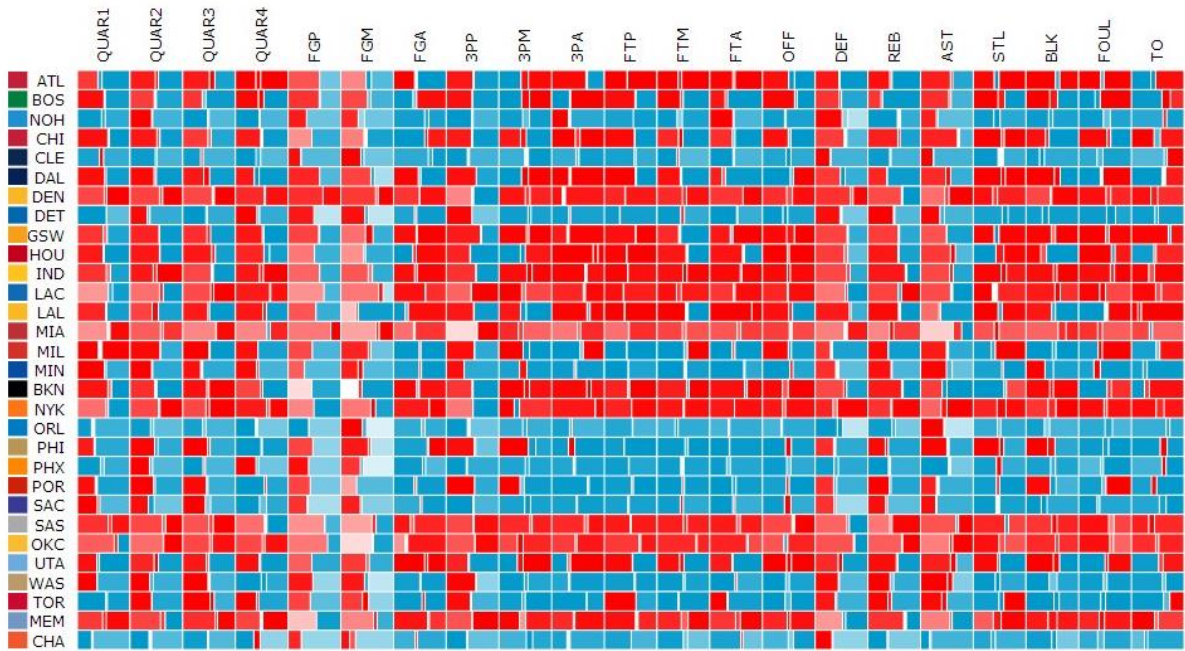
- 一开始是想通过一个球队一个赛季的八十二场比赛数据构建一个决定球队比赛胜负的决策树，构建决策树的算法中有用到“信息增益”等概念，由此想到了使用信息熵的一些方法计算球队获胜因素的相关性等。
- 每个维度（得分、失分、命中率、投篮总数等）是一个连续值，因此定义了针对连续值划分和与对手数据比较两种方式，目前只做了与对手数据比较的方法，连续值划分的方法下周再做。
- 一个球队的所有维度都是两两相关的，分别是球队自己的和对手的，如得分与对手得分，因此可以把维度缩减一半，并将连续值编码成离散值：数据比对手多标为 1，和对手相同标为 0，比对手少标为-1。把每个维度当作一个随机变量，可以计算出每个取值的概率。
- 计算每个维度与胜负维度的互信息，粗略得到各个维度与胜负维度的相关性，颜色越深表示相关性越高，如下图所示：

图 1：各个维度与胜负维度的互信息



从图 1 中可以看到整体上 FGP（投篮命中率），FGM（投篮命中数），DEF（防守篮板）和 AST（助攻总数）是影响胜负比较关键的维度，这非常符合现实情况。然而，其中也有些例外，如 DEN 这个球队，它的 FGP 与胜负维度相关性很小，而 3PP（三分命中率）与胜负的维度相关性则较高，这说明这是一支更侧重于三分的进攻性球队。当然，由于样本数据量非常小，上图展示出的准确性不一定高，所以我们需要更多的数据。

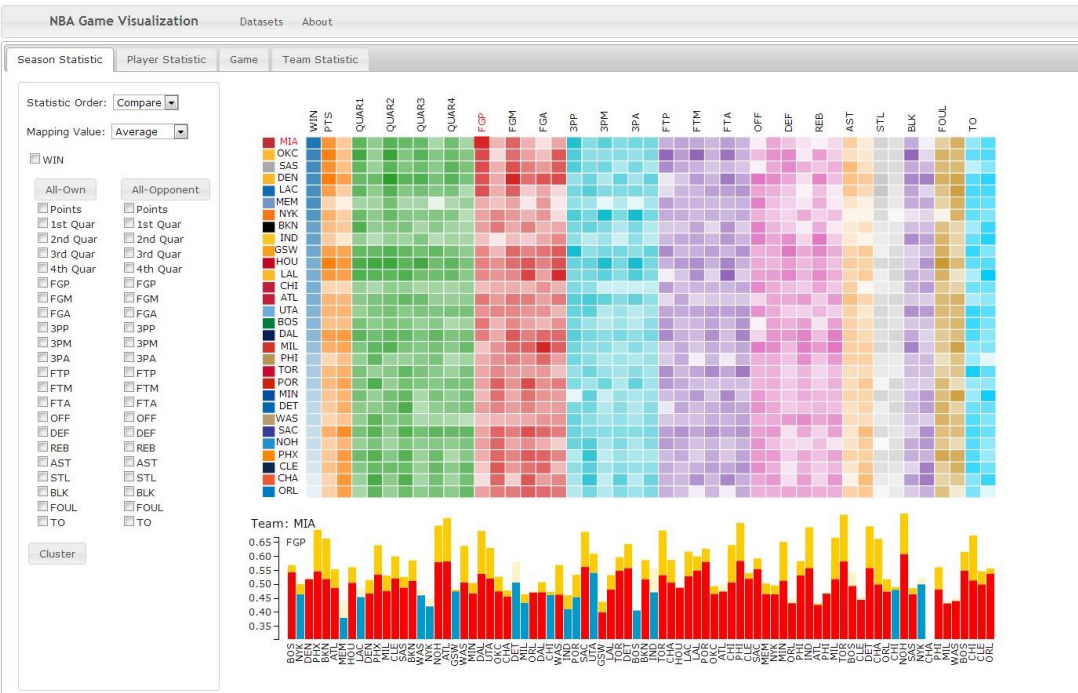
图 2：各个维度的分段熵可视化



上面这个略显杂乱的可视化的编码方式如下：如图所示，每个维度的长矩形里面划分了三个区间，从左到右分别是区间取值 1（该数据比对手多），0（该数据和对手相同），-1（该数据比对手少）；区间的长度编码了该取值在所有比赛中出现的概率；区间的颜色只用红色和蓝色编码，红色表示在该区间内，球队赢得比赛的概率较大，蓝色表示在该区间内球队输球的概率比较大；颜色的深浅（饱和度）则编码了该区间的熵，颜色越浅，熵越小，即稳定性越高。

因此，对与上图，我们的关注点应该是那些长度较长（发生概率较大）、颜色较小（熵小，稳定）的区间。另外，如果该维度下的三个区间颜色差不多，那么该维度对于该球队是可以忽略的。

- 这周也对视图和交互做了一些完善：



2. 找到一个数据非常完整的网站 basketball-reference.com, 这个网站的数据统计做的非常棒, 很多我意想不到的数据都有, 而且也做了一些非常简单的可视化。这周让洪瀚去把这些数据搞下来。
3. 读了一些课外书, 如《信号与噪声》、《大数据时代》, 其中有些主要观点如下:
 - 在大数据时代, 我们可以分析更多的数据, 有时候甚至可以处理和某个特别现象相关的所有数据, 而不再依赖于随机采用。
 - 研究数据如此之多, 以至于我们不再热衷于追求精确度。
 - 不再热衷与寻找因果关系。相关关系也许不能准确地告知我们某件事情为什么发生, 但是它会提醒我们这件事情正在发生。
 - 大数据的核心在与预测。

下周计划:

1. 得到比赛的文字直播数据, 加入系统中; 对更大量的数据做计算; 连续区间划分的计算。